Privacy Aware Learning

John C. Duchi[1] **Michael I. Jordan**[1,2] **Martin J. Wainwright**[1,2]
[1]Department of Electrical Engineering and Computer Science, [2]Department of Statistics
University of California, Berkeley
Berkeley, CA USA 94720
{jduchi,jordan,wainwrig}@eecs.berkeley.edu

Abstract

We study statistical risk minimization problems under a version of privacy in which the data is kept confidential even from the learner. In this local privacy framework, we establish sharp upper and lower bounds on the convergence rates of statistical estimation procedures. As a consequence, we exhibit a precise tradeoff between the amount of privacy the data preserves and the utility, measured by convergence rate, of any statistical estimator.

1 Introduction

There are natural tensions between learning and privacy that arise whenever a learner must aggregate data across multiple individuals. The learner wishes to make optimal use of each data point, but the providers of the data may wish to limit detailed exposure, either to the learner or to other individuals. It is of great interest to characterize such tensions in the form of quantitative tradeoffs that can be both part of the public discourse surrounding the design of systems that learn from data and can be employed as controllable degrees of freedom whenever such a system is deployed.

We approach this problem from the point of view of statistical decision theory. The decision-theoretic perspective offers a number of advantages. First, the use of loss functions and risk functions provides a compelling formal foundation for defining "learning," one that dates back to Wald [28] in the 1930's, and which has seen continued development in the context of research on machine learning over the past two decades. Second, by formulating the goals of a learning system in terms of loss functions, we make it possible for individuals to assess whether the goals of a learning system align with their own personal utility, and thereby determine the extent to which they are willing to sacrifice some privacy. Third, an appeal to decision theory permits abstraction over the details of specific learning procedures, providing (under certain conditions) minimax lower bounds that apply to any specific procedure. Finally, the use of loss functions, in particular convex loss functions, in the design of a learning system allows powerful tools of optimization theory to be brought to bear.

In more formal detail, our framework is as follows. Given a compact convex set $\Theta \subset \mathbb{R}^d$, we wish to find a parameter value $\theta \in \Theta$ achieving good average performance under a loss function $\ell : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}_+$. Here the value $\ell(X, \theta)$ measures the performance of the parameter vector $\theta \in \Theta$ on the sample $X \in \mathcal{X}$, and $\ell(x, \cdot) : \mathbb{R}^d \to \mathbb{R}_+$ is convex for $x \in \mathcal{X}$. We measure the expected performance of $\theta \in \Theta$ via the risk function

$$R(\theta) := \mathbb{E}[\ell(X, \theta)]. \tag{1}$$

In the standard formulation of statistical risk minimization, a method $\mathcal{M}$ is given n samples $X_1, \ldots, X_n$, and outputs an estimate θ_n approximately minimizing $R(\theta)$. Instead of allowing $\mathcal{M}$ access to the samples X_i, however, we study the effect of giving only a perturbed view Z_i of each datum X_i, quantifying the rate of convergence of $R(\theta_n)$ to $\inf_{\theta \in \Theta} R(\theta)$ as a function of both the number of samples n and the amount of privacy Z_i provides for X_i.

There is a long history of research at the intersection of privacy and statistics, where there is a natural competition between maintaining the privacy of elements in a dataset $\{X_1, \ldots, X_n\}$ and the output of statistical procedures. Study of this issue goes back at least to the 1960s, when Warner [29] suggested privacy-preserving methods for survey sampling. Recently, there has been substantial work on privacy—focusing on a measure known as differential privacy [12]—in statistics, computer science, and other fields. We cannot hope to do justice to the large body of related work, referring the reader to the survey by Dwork [10] and the statistical framework studied by Wasserman and Zhou [30] for background and references.

In this paper, we study *local privacy* [13, 17], in which each datum X_i is kept private from the method $\mathcal{M}$. The goal of many types of privacy is to guarantee that the output $\widehat{\theta}_n$ of the method $\mathcal{M}$ based on the data cannot be used to discover information about the individual samples $X_1, \ldots, X_n$, but locally private algorithms access only disguised views of each datum X_i. Local algorithms are among the most classical approaches to privacy, tracing back to Warner's work on randomized response [29], and rely on communication only of some disguised view Z_i of each true sample X_i. Locally private algorithms are natural when the providers of the data—the population sampled to give $X_1, \ldots, X_n$—do not trust even the statistician or statistical method $\mathcal{M}$, but the providers are interested in the parameters θ^* minimizing $R(\theta)$. For example, in medical applications, a participant may be embarrassed about his use of drugs, but if the loss ℓ is able to measure the likelihood of developing cancer, the participant has high utility for access to the optimal parameters θ^*. In essence, we would like the statistical procedure $\mathcal{M}$ to learn *from* the data $X_1, \ldots, X_n$ but not *about* it.

Our goal is to understand the fundamental tradeoffs between maintaining privacy while still retaining the utility of the statistical inference method $\mathcal{M}$. Though intuitively there must be some tradeoff, quantifying it precisely has been difficult. In the machine learning literature, Chaudhuri et al. [7] develop differentially private empirical risk minimization algorithms, and Dwork and Lei [11] and Smith [26] analyze similar statistical procedures, but do not show that there must be negative effects of privacy. Rubinstein et al. [24] are able to show that it is impossible to obtain a useful parameter vector θ that is substantially differentially private; it is unclear whether their guarantees are improvable. Recent work by Hall et al. [15] gives sharp minimax rates of convergence for differentially private histogram estimation. Blum et al. [5] also give lower bounds on the closeness of certain statistical quantities computed from the dataset, though their upper and lower bounds do not match. Sankar et al. [25] provide rate-distortion theorems for utility models involving information-theoretic quantities, which has some similarity to our risk-based framework, but it appears challenging to map their setting onto ours. The work most related to ours is probably that of Kasiviswanathan et al. [17], who show that that locally private algorithms coincide with concepts that can be learned with polynomial sample complexity in Kearns's statistical query (SQ) model. In contrast, our analysis addresses sharp rates of convergence, and applies to estimators for a broad class of convex risks (1).

2 Main results and approach

Our approach to local privacy is based on a worst-case measure of mutual information, where we view privacy preservation as a game between the providers of the data—who wish to preserve privacy—and nature. Recalling that the method sees only the perturbed version Z_i of X_i, we adopt a uniform variant of the mutual information $I(Z_i; X_i)$ between the random variables X_i and Z_i as our measure for privacy. This use of mutual information is by no means original [13, 25], but because standard mutual information has deficiencies as a measure of privacy [e.g. 13], we say the distribution Q generating Z from X is private only if $I(X; Z)$ is small for *all* possible distributions P on X (possibly subject to constraints). This is similar to the worst-case information approach of Evfimievski et al. [13], which limits privacy breaches. (In the long version of this paper [9] we also consider differentially private algorithms.)

The central consequences of our main results are, under standard conditions on the loss functions ℓ, sharp upper and lower bounds on the possible convergence rates for estimation procedures when we wish to guarantee a level of privacy $I(X_i; Z_i) \le I^*$. We show there are problem dependent constants $a(\Theta, \ell)$ and $b(\Theta, \ell)$ such that the rates of convergence of *all possible procedures* are lower bounded by $a(\Theta, \ell)/\sqrt{nI^*}$ and that *there exist* procedures achieving convergence rates of $b(\Theta, \ell)/\sqrt{nI^*}$, where the ratio $b(\Theta, \ell)/a(\Theta, \ell)$ is upper bounded by a universal constant. Thus, we establish and quantify explicitly the tradeoff between statistical estimation and the amount of privacy.

We show that stochastic gradient descent is one procedure that achieves the optimal convergence rates, which means additionally that our upper bounds apply in streaming and online settings, requiring only a fixed-size memory footprint. Our subsequent analysis builds on this favorable property of gradient-based methods, whence we focus on statistical estimation procedures that access data through the subgradients of the loss functions $\partial\ell(X, \theta)$. This is a natural restriction. Gradients of the loss ℓ are asymptotically sufficient [18] (in an asymptotic sense, gradients contain *all* of the statistical information for risk minimization problems), stochastic gradient-based estimation procedures are (sample) minimax optimal and Bahadur efficient [23, 1, 27, Chapter 8], many estimation procedures are gradient-based [20, 6], and distributed optimization procedures that send gradient information across a network to a centralized procedure $\mathcal{M}$ are natural [e.g. 3]. Our mechanism gives $\mathcal{M}$ access to a vector Z_i that is a stochastic (sub)gradient of the loss evaluated on the sample X_i at a parameter θ of the method's choosing:

$$\mathbb{E}[Z_i \mid X_i, \theta] \in \partial\ell(X_i, \theta), \tag{2}$$

where $\partial\ell(X_i, \theta)$ denotes the subgradient set of the function $\theta \mapsto \ell(X_i, \theta)$. In a sense, the unbiasedness of the subgradient inclusion (2) is information-theoretically necessary [1].

To obtain upper and lower bound on the convergence rate of estimation procedures, we provide a two-part analysis. One part requires studying saddle points of the mutual information $I(X; Z)$ (as a function of the distributions P of X and $Q(\cdot \mid X)$ of Z) under natural constraints that allow inference of the optimal parameters θ^* for the risk R. We show that for certain classes of loss functions ℓ and constraints on the communicated version Z_i of the data X_i, there is a unique distribution $Q(\cdot \mid X_i)$ that attains the smallest possible mutual information $I(X; Z)$ for all distributions on X. Using this unique distribution, we can adapt information-theoretic techniques for obtaining lower bounds on estimation [31, 1] to derive our lower bounds. The uniqueness results for the conditional distribution Q show that no algorithm guaranteeing privacy between $\mathcal{M}$ and the samples X_i can do better. We can obtain matching upper bounds by application of known convergence rates for stochastic gradient and mirror descent algorithms [20, 21], which are computationally efficient.

3 Optimal learning rates and tradeoffs

Having outlined our general approach, we turn in this section to providing statements of our main results. Before doing so, we require some formalization of our notions of privacy and error measures, which we now provide.

3.1 Optimal Local Privacy

We begin by describing in slightly more detail the communication protocol by which information about the random variables X is communicated to the procedure $\mathcal{M}$. We assume throughout that there exist two d-dimensional compact sets C, D, where $C \subset \operatorname{int} D \subset \mathbb{R}^d$, and we have that $\partial\ell(x, \theta) \subset C$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. We wish to maximally "disguise" the random variable X with the random variable Z satisfying $Z \in D$. Such a setting is natural; indeed, many online optimization and stochastic approximation algorithms [34, 21, 1] assume that for any $x \in \mathcal{X}$ and $\theta \in \Theta$, if $g \in \partial\ell(x, \theta)$ then $\|g\| \le L$ for some norm $\|\cdot\|$. We may obtain privacy by allowing a perturbation to the subgradient g so long as the perturbation lives in a (larger) norm ball of radius $M \ge L$, so that $C = \{g \in \mathbb{R}^d : \|g\| \le L\} \subset D = \{g \in \mathbb{R}^d : \|g\| \le M\}$.

Now let X have distribution P, and for each $x \in \mathcal{X}$, let $Q(\cdot \mid x)$ denote the regular conditional probability measure of Z given that $X = x$. Let $Q(\cdot)$ denote the marginal probability defined by $Q(A) = \mathbb{E}_P[Q(A \mid X)]$. The mutual information between X and Z is the expected Kullback-Leibler (KL) divergence between $Q(\cdot \mid X)$ and $Q(\cdot)$:

$$I(P, Q) = I(X; Z) := \mathbb{E}_P \left[D_{\rm kl} \left(Q(\cdot \mid X) || Q(\cdot) \right) \right]. \tag{3}$$

We view the problem of privacy as a game between the adversary controlling P and the data owners, who use Q to obscure the samples X. In particular, we say a distribution Q guarantees a level of privacy I^* if and only if $\sup_P I(P, Q) \le I^*$. (Evfimievski et al. [13, Definition 6] present a similar condition.) Thus we seek a saddle point P^*, Q^* such that

$$\sup_P I(P, Q^*) \le I(P^*, Q^*) \le \inf_Q I(P^*, Q), \tag{4}$$

where the first supremum is taken over all distributions P on X such that $\nabla \ell(X, \theta) \in C$ with P-probability 1, and the infimum is taken over all regular conditional distributions Q such that if $Z \sim Q(\cdot \mid X)$, then $Z \in D$ and $\mathbb{E}_Q[Z \mid X, \theta] = \nabla \ell(X, \theta)$. Indeed, if we can find P^* and Q^* satisfying the saddle point (4), then the trivial direction of the max-min inequality yields

$$\sup_P \inf_Q I(P, Q) = I(P^*, Q^*) = \inf_Q \sup_P I(P, Q).$$

To fully formalize this idea and our notions of privacy, we define two collections of probability measures and associated losses. For sets $C \subset D \subset \mathbb{R}^d$, we define the source set

$$\mathcal{P}(C) := \{\text{Distributions } P \text{ such that } \operatorname{supp} P \subset C\} \tag{5a}$$

and the set of regular conditional distributions (r.c.d.'s), or communicating distributions,

$$\mathcal{Q}(C, D) := \left\{\text{r.c.d.'s } Q \text{ s.t. } \operatorname{supp} Q(\cdot \mid c) \subset D \text{ and } \int_D z dQ(z \mid c) = c \text{ for } c \in C\right\}. \tag{5b}$$

The definitions (5a) and (5b) formally define the sets over which we may take infima and suprema in the saddle point calculations, and they capture what may be communicated. The conditional distributions $Q \in \mathcal{Q}(C, D)$ are defined so that if $\nabla \ell(x, \theta) \in C$ then $\mathbb{E}_Q[Z \mid X, \theta] := \int_D z dQ(z \mid \nabla \ell(x, \theta)) = \nabla \ell(x, \theta)$. We now make the following key definition:

Definition 1. *The conditional distribution Q^* satisfies* optimal local privacy *for the sets $C \subset D \subset \mathbb{R}^d$ at level I^* if*

$$\sup_P I(P, Q^*) = \inf_Q \sup_P I(P, Q) = I^*,$$

where the supremum is taken over distributions $P \in \mathcal{P}(C)$ and the infimum is taken over regular conditional distributions $Q \in \mathcal{Q}(C, D)$.

If a distribution Q^* satisfies optimal local privacy, then it guarantees that even for the worst possible distribution on X, the information communicated about X is limited. In a sense, Definition 1 captures the natural competition between privacy and learnability. The method $\mathcal{M}$ specifies the set D to which the data Z it receives must belong; the "teachers," or owners of the data X, choose the distribution Q to guarantee as much privacy as possible subject to this constraint. Using this mechanism, if we can characterize a unique distribution Q^* attaining the infimum (4) for P^* (and by extension, for any P), then we may study the effects of privacy between the method $\mathcal{M}$ and X.

3.2 Minimax error and loss functions

Having defined our privacy metric, we now turn to our original goal: quantification of the effect privacy has on statistical estimation rates. Let $\mathcal{M}$ denote any statistical procedure or method (that uses n stochastic gradient samples) and let θ_n denote the output of $\mathcal{M}$ after receiving n such samples. Let P denote the distribution according to which samples X are drawn. We define the (random) error of the method $\mathcal{M}$ on the risk $R(\theta) = \mathbb{E}[\ell(X, \theta)]$ after receiving n sample gradients as

$$\epsilon_n(\mathcal{M}, \ell, \Theta, P) := R(\theta_n) - \inf_{\theta \in \Theta} R(\theta) = \mathbb{E}_P[\ell(X, \theta_n)] - \inf_{\theta \in \Theta} \mathbb{E}_P[\ell(X, \theta)]. \tag{6}$$

In our settings, in addition to the randomness in the sampling distribution P, there is additional randomness from the perturbation applied to stochastic gradients of the objective $\ell(X, \cdot)$ to mask X from the statistitician. Let Q denote the regular conditional probability—the channel distribution—whose conditional part is defined on the range of the subgradient mapping $\partial \ell(X, \cdot)$. As the output θ_n of the statistical procedure $\mathcal{M}$ is a random function of both P and Q, we measure the expected sub-optimality of the risk according to both P and Q. Now, let $\mathfrak{L}$ be a collection of loss functions, where $\mathfrak{L}(P)$ denotes the losses $\ell : \operatorname{supp} P \times \Theta \to \mathbb{R}$ belonging to $\mathfrak{L}$. We define the minimax error

$$\epsilon_n^*(\mathfrak{L}, \Theta) := \inf_{\mathcal{M}} \sup_{\ell \in \mathfrak{L}(P), P} \mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)], \tag{7}$$

where the expectation is taken over the random samples $X \sim P$ and $Z \sim Q(\cdot \mid X)$. We characterize the minimax error (7) for several classes of loss functions $\mathfrak{L}(P)$, giving sharp results when the privacy distribution Q satisfies optimal local privacy.

We assume that our collection of loss functions obey certain natural smoothness conditions, which are often (as we see presently) satisfied. We define the class of losses as follows.

Definition 2. *Let $L > 0$ and $p \geq 1$. The set of* (L, p)-loss functions *are those measurable functions $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $x \in \mathcal{X}$, the function $\theta \mapsto \ell(x, \theta)$ is convex and*

$$|\ell(x, \theta) - \ell(x, \theta')| \leq L \left\|\theta - \theta'\right\|_q \tag{8}$$

for any $\theta, \theta' \in \Theta$, where q is the conjugate of p: $1/p + 1/q = 1$.

A loss ℓ satisfies the condition (8) if and only if for all $\theta \in \Theta$, we have the inequality $\left\|g\right\|_p \leq L$ for any subgradient $g \in \partial \ell(x, \theta)$ (e.g. [16]). We give a few standard examples of such loss functions. First, we consider finding a multi-dimensional median, in which case the data $x \in \mathbb{R}^d$ and $\ell(x, \theta) = L \left\|\theta - x\right\|_1$. This loss is L-Lipschitz with respect to the ℓ_1 norm, so it belongs to the class of (L, ∞) losses. A second example includes classification problems, using either the hinge loss or logistic regression loss. In these cases, the data comes in pairs $x = (a, b)$, where $a \in \mathbb{R}^d$ is the set of regressors and $b \in \{-1, 1\}$ is the label; the losses are

$$\ell(x, \theta) = \left[1 - b \left\langle a, \theta \right\rangle\right]_+ \quad \text{or} \quad \ell(x, \theta) = \log\left(1 + \exp(-b \left\langle a, \theta \right\rangle)\right)$$

By computing (sub)gradients, we may verify that each of these belong to the class of (L, p)-losses if and only if the data a satisfies $\left\|a\right\|_p \leq L$, which is a common assumption [7, 24].

The privacy-guaranteeing channel distributions Q^* we construct in Section 4 are motivated by our concern with the (L, p) families of loss functions. In our model of computation, the learning method $\mathcal{M}$ queries the loss $\ell(X_i, \cdot)$ at the point θ; the owner of the datum X_i then computes the subgradient $\partial \ell(X_i, \theta)$ and returns a masked version Z_i with the property that $\mathbb{E}[Z_i \mid X_i, \theta] \in \partial \ell(X_i, \theta)$. In the following two theorems, we give lower bounds on ϵ_n^* for the (L, ∞) and $(L, 1)$ families of loss functions under the constraint that the channel distribution Q *must* guarantee that a limited amount of information $I(X_i; Z_i)$ is communicated: the channel distribution Q satisfies our Definition 1 of optimal local privacy.

3.3 Main theorems

We now state our two main theorems, deferring proofs to Appendix B. Our first theorem applies to the class of (L, ∞) loss functions (recall Definition 2). We assume that the set to which the perturbed data Z must belong is $[-M_\infty, M_\infty]^d$, where $M_\infty \geq L$. We state two variants of the theorem, as one gives sharper results for an important special case.

Theorem 1. *Let $\mathfrak{L}$ be the collection of (L, ∞) loss functions and assume the conditions of the preceding paragraph. Let Q satisfy be optimally private for the collection $\mathfrak{L}$. Then*

(a) If Θ contains the ℓ_∞ ball of radius r,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \cdot \frac{M_\infty r d}{\sqrt{n}}.$$

(b) If $\Theta = \{\theta \in \mathbb{R}^d : \left\|\theta\right\|_1 \leq r\}$,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{r M_\infty \sqrt{\log(2d)}}{17\sqrt{n}}.$$

For our second theorem, we assume that the loss functions $\mathfrak{L}$ consist of $(L, 1)$ losses, and that the perturbed data must belong to the ℓ_1 ball of radius M_1, i.e., $Z \in \{z \in \mathbb{R}^d \mid \left\|z\right\|_1 \leq M_1\}$. Setting $M = M_1 / L$, we define (these constants relate to the optimal local privacy distribution for ℓ_1-balls)

$$\gamma := \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2 - 1)}}{2(M-1)}\right), \quad \text{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)}. \tag{9}$$

Theorem 2. *Let $\mathfrak{L}$ be the collection of $(L, 1)$ loss functions and assume the conditions of the preceding paragraph. Let Q be optimally locally private for the collection $\mathfrak{L}$. Then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \cdot \frac{r L \sqrt{d}}{\sqrt{n} \Delta(\gamma)}.$$

Remarks We make two main remarks about Theorems 1 and 2. First, we note that each result yields a minimax rate for stochastic optimization problems when there is no random distribution Q. Indeed, in Theorem 1, we may take $M_\infty = L$, in which case (focusing on the second statement of the theorem) we obtain the lower bound $rL\sqrt{\log(2d)}/17\sqrt{n}$ when $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$. Mirror descent algorithms [20, 21] attain a matching upper bound (see the long version of this paper [9, Sec. 3.3] for more substantial explanation). Moreover, our analysis is sharper than previous analyses [1, 20], as none (to our knowledge) recover the logarithmic dependence on the dimension d, which is evidently necessary. Theorem 2 provides a similar result when we take $M_1 \downarrow L$, though in this case stochastic gradient descent attains the matching upper bound.

Our second set of remarks are somewhat more striking. In these, we show that the lower bounds in Theorems 1 and 2 give sharp tradeoffs between the statistical rate of convergence for any statistical procedure and the desired privacy of a user. We present two corollaries establishing this tradeoff. In each corollary, we look ahead to Section 4 and use one of Propositions 1 or 2 to derive a bijection between the size M_∞ or M_1 of the perturbation set and the amount of privacy—as measured by the worst case mutual information I^*—provided. We then combine Theorems 1 and 2 with results on stochastic approximation to demonstrate the tradeoffs.

Corollary 1. *Let the conditions of Theorem 1(b) hold, and assume that $M_\infty \ge 2L$. Assume Q^* satisfies optimal local privacy at information level I^*. For universal constants $c \le C$,*

$$c \cdot \frac{rL\sqrt{d \log d}}{\sqrt{nI^*}} \le \epsilon_n^*(\mathfrak{L}, \Theta) \le C \cdot \frac{rL\sqrt{d \log d}}{\sqrt{nI^*}}.$$

Proof Since $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$, mirror descent [2, 21, 20, Chapter 5], using n unbiased stochastic gradient samples whose ℓ_∞ norms are bounded by M_∞, obtains convergence rate $\mathcal{O}(M_\infty r\sqrt{\log d}/\sqrt{n})$. This matches the second statement of Theorem 1. Now fix our desired amount of mutual information I^*. From the remarks following Proposition 1, if we must guarantee that $I^* \ge \sup_P I(P, Q)$ for any distribution P and loss function ℓ whose gradients are bounded in ℓ_∞-norm by L, we *must* (by the remarks following Proposition 1) have

$$I^* \asymp \frac{dL^2}{M_\infty^2}.$$

Up to higher-order terms, to guarantee a level of privacy with mutual information I^*, we must allow gradient noise up to $M_\infty = L\sqrt{d/I^*}$. Using the bijection between M_∞ and the maximal allowed mutual information I^* under local privacy that we have shown, we substitute $M_\infty = L\sqrt{d}/\sqrt{I^*}$ into the upper and lower bounds that we have already attained. □

Similar upper and lower bounds can be obtained under the conditions of part (a) of Theorem 1, where we need not assume Θ is an ℓ_1-ball, but we lose a factor of $\sqrt{\log d}$ in the lower bound. Now we turn to a parallel result, but applying Theorem 2 and Proposition 2.

Corollary 2. *Let the conditions of Theorem 2 hold and assume that $M_1 \ge 2L$. Assume that Q^* satisfies optimal local privacy at information level I^*. For universal constants $c \le C$,*

$$c \cdot \frac{rLd}{\sqrt{nI^*}} \le \epsilon_n^*(\mathfrak{L}, \Theta) \le C \cdot \frac{rLd}{\sqrt{nI^*}}.$$

Proof By the conditions of optimal local privacy (Proposition 2 and Corollary 3), to have $I^* \ge \sup_P I(P, Q)$ for any loss ℓ whose gradients are bounded in ℓ_1-norm by L, we must have

$$I^* \asymp \frac{dL^2}{2M_1^2},$$

using Corollary 3. Rewriting this, we see that we must have $M_1 = L\sqrt{d/2I^*}$ (to higher-order terms) to be able to guarantee an amount of privacy I^*. As in the ℓ_∞ case, we have a bijection between the multiplier M_1 and the amount of information I^* and can apply similar techniques. Indeed, stochastic gradient descent (SGD) enjoys the following convergence guarantees (e.g. [21]). Let $\Theta \subseteq \mathbb{R}^d$ be contained in the ℓ_∞ ball of radius r and the gradients of the loss ℓ belong to the ℓ_1-ball of radius M_1. Then SGD has $\epsilon_n^*(\mathfrak{L}, \Theta) \le CM_1 r\sqrt{d}/\sqrt{n}$. Now apply the lower bound provided by Theorem 2 and substitute for M_1. □

4 Saddle points, optimal privacy, and mutual information

In this section, we explore conditions for a distribution Q^* to satisfy optimal local privacy, as given by Definition 1. We give characterizations of necessary and sufficient conditions based on the compact sets $C \subset D$ for distributions P^* and Q^* to achieve the saddle point (4). Our results can be viewed as rate distortion theorems [14, 8] (with source P and channel Q) for certain compact alphabets, though as far as we know, they are all new. Thus we sometimes refer to the conditional distribution Q, which is designed to maintain the privacy of the data X by communication of Z, as the channel distribution. Since we wish to bound $I(X; Z)$ for arbitrary losses ℓ, we must address the case when $\ell(X, \theta) = \langle \theta, X \rangle$, in which case $\nabla \ell(X, \theta) = X$; by the data-processing inequality [14, Chapter 5] it is thus no loss of generality to assume that $X \in C$ and that $\mathbb{E}[Z \mid X] = X$.

We begin by defining the types of sets C and D that we use in our characterization of privacy. As we see in Section 3, such sets are reasonable for many applications. We focus on the case when the compact sets C and D are (suitably symmetric) norm balls:

Definition 3. *Let $C \subset \mathbb{R}^d$ be a compact convex set with extreme points $u_i \in \mathbb{R}^d$, $i \in I$ for some index set I. Then C is* rotationally invariant through its extreme points *if $\|u_i\|_2 = \|u_j\|_2$ for each i, j, and for any unitary matrix U such that $Uu_i = u_j$ for some $i \neq j$, then $UC = C$.*

Some examples of convex sets rotationally invariant through their extreme points include ℓ_p-norm balls for $p = 1, 2, \infty$, though ℓ_p-balls for $p \notin \{1, 2, \infty\}$ are not. The following theorem gives a general characterization of the minimax mutual information for rotationally invariant norm balls with finite numbers of extreme points by providing saddle point distributions P^* and Q^*. We provide the proof of Theorem 3 in Section A.1.

Theorem 3. *Let C be a compact, convex, polytope rotationally invariant through its extreme points $\{u_i\}_{i=1}^m$ and $D = (1 + \alpha)C$ for some $\alpha > 0$. Let Q^* be the conditional distribution on $Z \mid X$ that maximizes the entropy $H(Z \mid X = x)$ subject to the constraints that*

$$\mathbb{E}_Q[Z \mid X = x] = x$$

for $x \in C$ and that Z is supported on $(1 + \alpha)u_i$ for $i = 1, \ldots, m$. Then Q^* satisfies Definition 1, optimal local privacy, and Q^* is (up to measure zero sets) unique. Moreover, the distribution P^* uniform on $\{u_i\}_{i=1}^m$ uniquely attains the saddle point (4)*.*

Remarks: While in the theorem we assume that $Q^*(\cdot \mid X = x)$ maximizes the entropy for each $x \in C$, this is not in fact essential. In fact, we may introduce a random variable X' between X and Z: let X' be distributed among the extreme points $\{u_i\}_{i=1}^m$ of C in *any* way such that $\mathbb{E}[X' \mid X] = X$, then use the maximum entropy distribution $Q^*(\cdot \mid u_i)$ defined in the theorem when $X \in \{u_i\}_{i=1}^m$ to sample Z from X'. The information processing inequality [14, Chapter 5] guarantees the Markov chain $X \to X' \to Z$ satisfies the minimax bound $I(X; Z) \leq \inf_Q \sup_P I(P, Q)$.

With Theorem 3 in place, we can explicitly characterize the distributions achieving optimal local privacy (recall Definition 1) for ℓ_1 and ℓ_∞ balls. We present the propositions in turn, providing some discussion here and deferring proofs to Appendices A.2 and A.3.

First, consider the case where $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$. For notational convenience, we define the binary entropy $h(p) = -p \log p - (1 - p) \log(1 - p)$. We have

Proposition 1. *Let $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$ be random variables with $M \geq 1$ and $\mathbb{E}[Z \mid X] = X$ almost surely. Define Q^* to be the conditional distribution on $Z \mid X$ such that the coordinates of Z are independent, have range $\{-M, M\}$, and*

$$Q^*(Z_i = M \mid X) = \frac{1}{2} + \frac{X_i}{2M} \quad \text{and} \quad Q^*(Z_i = -M \mid X) = \frac{1}{2} - \frac{X_i}{2M}.$$

Then Q^* satisfies Definition 1, optimal local privacy, and moreover,

$$\sup_P I(P, Q^*) = d - d \cdot h\left(\frac{1}{2} + \frac{1}{2M}\right).$$

Before continuing, we give a more intuitive understanding of Proposition 1. Concavity implies that for $a, b > 0$, $\log(a) \leq \log b + b^{-1}(a - b)$, or $-\log(a) \geq -\log(b) + b^{-1}(b - a)$, so in particular

$$h\left(\frac{1}{2} + \frac{1}{2M}\right) \geq -\left(\frac{1}{2} + \frac{1}{2M}\right)\left(-\log 2 - \frac{1}{M}\right) - \left(\frac{1}{2} - \frac{1}{2M}\right)\left(-\log 2 + \frac{1}{M}\right) = \log 2 - \frac{1}{M^2}.$$

That is, we have for *any* distribution P on $X \in [-1,1]^d$ that (in natural logarithms)

$$I(P, Q^*) \le \frac{d}{M^2} \quad \text{and} \quad I(P, Q^*) = \frac{d}{M^2} + \mathcal{O}(M^{-3}).$$

We now consider the case when $X \in \left\{x \in \mathbb{R}^d \mid \|x\|_1 \le 1\right\}$ and $Z \in \left\{z \in \mathbb{R}^d \mid \|z\|_1 \le M\right\}$. Here the arguments are slightly more complicated, as the coordinates of the random variables are no longer independent, but Theorem 3 still allows us to explicitly characterize the saddle point of the mutual information.

Proposition 2. *Let $X \in \{x \in \mathbb{R}^d \mid \|x\|_1 \le 1\}$ and $Z \in \{z \in \mathbb{R}^d \mid \|z\|_1 \le M\}$ be random variables with $M > 1$. Define the parameter γ as in Eq.* (9)*, and let Q^* be the distribution on $Z \mid X$ such that Z is supported on $\{\pm M e_i\}_{i=1}^d$, and*

$$Q^*(Z = Me_i \mid X = e_i) = \frac{e^\gamma}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{10a}$$

$$Q^*(Z = -Me_i \mid X = e_i) = \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{10b}$$

$$Q^*(Z = \pm Me_j \mid X = e_i, j \ne i) = \frac{1}{e^\gamma + e^{-\gamma} + (2d-2)}. \tag{10c}$$

(For $X \notin \{\pm e_i\}$, define X' to be randomly selected in any way from among $\{\pm e_i\}$ such that $\mathbb{E}[X' \mid X] = X$, then sample Z conditioned on X' according to (10a)–(10c)*.) Then Q^* satisfies Definition 1, optimal local privacy, and*

$$\sup_P I(P, Q^*) = \log(2d) - \log\left(e^\gamma + e^{-\gamma} + 2d - 2\right) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2}.$$

We remark that the additional sampling to guarantee that $X' \in \{\pm e_i\}$ (where the conditional distribution Q^* is defined) can be accomplished simply: define the random variable X' so that $X' = e_i \operatorname{sign}(x_i)$ with probability $|x_i| / \|x\|_1$. Evidently $\mathbb{E}[X' \mid X] = x$, and $X \to X' \to Z$ for Z distributed according to Q^* defines a Markov chain as in our remarks following Theorem 3. Additionally, an asymptotic expansion allows us to gain a somewhat clearer picture of the values of the mutual information, though we do not derive upper bounds as we did for Proposition 1. We have the following corollary, proved in Appendix E.1.

Corollary 3. *Let Q^* denote the conditional distribution in Proposition 2. Then*

$$\sup_P I(P, Q^*) = \frac{d}{2M^2} + \Theta\left(\min\left\{\frac{d^3}{M^4}, \frac{\log^4(d)}{d}\right\}\right).$$

5 Discussion and open questions

This study leaves a number open issues and areas for future work. We study procedures that access each datum only once and through a perturbed view Z_i of the subgradient $\partial \ell(X_i, \theta)$, which allows us to use (essentially) any convex loss. A natural question is whether there are restrictions on the loss function so that a transformed version $(Z_1, \ldots, Z_n)$ of the data are sufficient for inference. Zhou et al. [33] study one such procedure, and nonparametric data releases, such as those Hall et al. [15] study, may also provide insights. Unfortunately, these (and other) current approaches require the data be aggregated by a trusted curator. Our constraints on the privacy-inducing channel distribution Q require that its support lie in some compact set. We find this restriction useful, but perhaps it possible to achieve faster estimation rates under other conditions. A better understanding of general privacy-preserving channels Q for alternative constraints to those we have proposed is also desirable.

These questions do not appear to have easy answers, especially when we wish to allow each provider of a single datum to be able to guarantee his or her own privacy. Nevertheless, we hope that our view of privacy and the techniques we have developed herein prove fruitful, and we hope to investigate some of the above issues in future work.

Acknowledgments We thank Cynthia Dwork, Guy Rothblum, and Kunal Talwar for feedback on early versions of this work. This material supported in part by ONR MURI grant N00014-11-1-0688 and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-11-1-0391. JCD was partially supported by an NDSEG fellowship and a Facebook fellowship.

References

[1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Trans. on Information Theory*, 58(5):3235–3249, 2012.

[2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

[4] P. Billingsley. *Probability and Measure*. Wiley, Second edition, 1986.

[5] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing*, 2008.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] K. Chaudhuri, C. Moneleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[9] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. URL `http://arxiv.org/abs/1210.2085`, 2012.

[10] C. Dwork. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, p. 1–19. Springer, 2008.

[11] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Fourty-First Annual ACM Symposium on the Theory of Computing*, 2009.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, p. 265–284, 2006.

[13] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, p. 211–222, 2003.

[14] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[15] R. Hall, A. Rinaldo, and L. Wasserman. Random differential privacy. URL `http://arxiv.org/abs/1112.2680`, 2011.

[16] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1996.

[17] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[18] L. Le Cam. On the asymptotic theory of estimation and hypothesis testing. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, p. 129–156, 1956.

[19] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):38–53, 1973.

[20] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[21] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[22] R. R. Phelps. *Lectures on Choquet's Theorem, Second Edition*. Springer, 2001.

[23] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[24] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[25] L. Sankar, S. R. Rajagopalan, and H. V. Poor. An information-theoretic approach to privacy. In *The 48th Allerton Conference on Communication, Control, and Computing*, p. 1220–1227, 2010.

[26] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing*, 2011.

[27] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

[28] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[29] S. L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[30] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[31] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[32] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, p. 423–435. Springer-Verlag, 1997.

[33] S. Zhou, J. Lafferty, and L. Wasserman. Compressed regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.

[34] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.